D300–D302 *Nucleic Acids Research, 2004, Vol. 32, Database issue*
DOI: 10.1093/nar/gkh087

GenProtEC: an updated and improved analysis of functions of *Escherichia coli* K-12 proteins

Margrethe H. Serres, Sulip Goswami and Monica Riley*

Marine Biological Laboratory, Woods Hole, MA 02540, USA

Received September 12, 2003; Revised and Accepted October 6, 2003

ABSTRACT

Using more than one approach to characterizing functions of unknown proteins, we now present in GenProtEC (http://genprotec.mbl.edu/) some level of function information for 87% of *Escherichia coli* K-12 proteins. A new approach that has yielded new information entails assigning content of structural domains and their functions to *E.coli* proteins. In addition, some earlier methods have been further refined to provide more meaningful data. The process of identifying and separating multimodular or fused proteins into component modules has been improved. As a result, groups of sequence-similar (paralogous) proteins have been refined. Experimental information from recent literature on previously unknown genes has been incorporated. We now use a rich system of characterizing cell roles which accents the fact that many proteins play more than one cellular role and therefore carry more than one designation from our detailed catalog of roles, MultiFun.

INTRODUCTION

Since GenProtEC (Genes and Proteins of *Escherichia coli*) was launched initially in 1995 its main goal has been to provide descriptions of functions for gene products encoded by the *E.coli* K-12 (strain MG1655) chromosome (1). GenProtEC was recently rebuilt as a MySQL-based database and new features have been added that supplement the approach of defining functions through sequence analysis of the encoded proteins. The database contains 4400 gene products of which 116 are RNA molecules and 4284 proteins, using the coding DNA sequences (CDSs) from the GenBank Accession No. U00096 with updates (2) (G. Plunkett, III, personal communication). Each gene product is represented by a gene page, which can be queried through various identifiers, including gene name, Blattner number (bnumber), Swiss-Prot ID, Enzyme Commission (EC) number(s) and gene product description. If more than one gene product matches the query statement, i.e. multiple enzymes with the same EC number, a short list is generated which the user can select from. The gene page gives function assignments for the gene product. The basis for the function description, i.e. experimental, phenotype, sequence similarity, similarity to structural domain(s), membership in a biochemically related group, is given in addition to links to published references. Also provided are gene synonym(s) and identifiers, gene type (such as enzyme, transport protein, regulatory protein, membrane protein) and EC number(s) for enzymes. Cellular role assignment(s) according to the MultiFun classification system (3) are also given for many of the gene products.

Genome-wide information is given on other accessible pages. The information on these pages is as follows:

(i) distribution of gene product types in *E.coli*;

(ii) multimodular proteins containing two or more independently encoded functions;

(iii) groups of sequence similar proteins (paralogs) and their group functions;

(iv) proteins grouped by biochemical and structural similarity;

(v) the MultiFun classification system for cellular roles;

(vi) a table of the most frequently found structural domains as classified by SCOP (4) present in *E.coli*.

The distribution of sources of knowledge about *E.coli* gene products is shown in Table 1. Sequence similarity continues to provide the greatest amount of non-experimental information, with domain structures and protein families making unique and valuable contributions. Table 2 shows the distribution of the types of gene products. Enzymes, including experimentally determined and putative assignments make up over one-third of the chromosomally encoded gene products. For 86.9% of the proteins there is some information on the functional characteristics of the proteins, while for 13.1% still none. New information includes functions inferred from structural domains present in previously unknown gene products. There are presently 3395 literature references linked to 2410 of the gene products. New functions continue to be discovered for *E.coli* gene products by experimentation. Since our 2001 published functional update (2) we have entered 106 new experimentally based function assignments to our database. One or more cellular role assignments have been made to 3344 of the *E.coli* gene products.

NEW FEATURES TO IMPROVE AND EXPAND FUNCTIONAL ANNOTATION

Structural domains

Structurally based domains in *E.coli* proteins have been identified in the Superfamily library of the SCOP database (4,5). These domains represent elements of known function such as binding sites, catalytic sites and more. We have made use of these domain assignments to enrich the annotation for

*To whom correspondence should be addressed. Tel: +1 508 289 7612; Fax: +1 508 457 4727; Email: mriley@mbl.edu

Table 1. Sources of knowledge about *E.coli* K-12 gene products

Degree of knowledge	Percent of gene products[a]
Experimental	53
Putative function (sequence similarity)	27
Phenotype	3
Structural domain(s)	52
Membership of paralogous protein family	44

[a]The sum is not 100% due to overlap in categories listed.

Table 2. Distribution of types of *E.coli* gene products

Gene type	Experimental	By similarity	Total (%)
Enzymes	1067	496	1563 (35.5)
Transporters	330	264	594 (13.5)
Regulators	223	144	367 (8.3)
Membrane proteins	56	183	239 (5.4)
Structural proteins	88	36	124 (2.8)
Protein factors	110	33	143 (3.3)
Carrier proteins	33	28	61 (1.4)
Leader peptides	12		12 (0.3)
RNAs	116		116 (2.6)
External origin	312		312 (7.1)
Phenotypes[a]			293 (6.7)
Unknown function			578 (13.1)
Total	2347	1184	4400

[a]Includes gene products with domain function assignments.

the *E.coli* proteins whose function is either not known (currently annotated as conserved proteins, conserved hypothetical proteins or unknown CDSs) or is predicted by sequence similarity (putative). Unknowns with no function assignments were reduced from 760 (17.3%) to 578 (13.1%). Although domain functions do not represent the entire function of the gene product, they can give important clues about the activity of the gene product. Domain function descriptions have also been added to 687 of the gene products with putative function assignments to enrich their annotation.

Protein families related by biochemistry and structure

We have identified members in *E.coli* of some well-known biochemically related families of proteins. Each family contains members that have structurally related binding and catalytic sites (known or imputed) and by biochemical mechanism of action. Some of the proteins in the groups are only distantly related when assessed by sequence criteria alone. Specifically, we have studied four families in detail: pyridoxal phosphate-dependent aminotransferases, thiamine-diphosphate (TPP)-dependent decarboxylases, crotonases and ATP-dependent CoA ligases.

REVISION OF EARLIER ANALYSES

Identification of multifunctional proteins

Some proteins are multifunctional, presumably a consequence of gene fusion, and represent two or more separately encoded functions. Gene fusion events evidently are dynamic and evolutionarily recent as the fusion partners differ from one bacterial species to another. To identify the fusion of proteins with truly independent origins, one needs to distinguish complete protein functions from large functional domains. We have expanded the data set used to identify fused proteins from sequence alignments by including alignments with polypeptides encoded by 49 additional genomes. The sequence alignments were done by Darwin (6) with the requirement that an alignment be at least 83 amino acids long and have a PAM (point accepted mutation) distance of 200 or less. *E.coli* sequences that aligned to the full length of more than one single function polypeptide (either paralogous or orthologous) were split into modules as shown in Figure 1. Currently we have identified 101 fused proteins with two independent functions and seven proteins with three independent functions. The average length of the fused proteins is 635 amino acids and the average length of their encoded modules is 300 amino acids. The module size is more comparable to the average of 309 amino acids for the remaining (non-fused) or unimodular proteins.

In GenProtEC fused genes are presented in fused form and are annotated with their complete set of functions. In addition, independent functions are listed for the module components. For ThrA, as an example, the complete function is bifunctional: aspartokinase I (N-teminal); homoserine dehydrogenase I (C-terminal). The two modules are annotated separately as b0002_1 and b0002_2, with the two enzymes and EC numbers separately attributed. MultiFun assignment is perforce assigned at the level of module.

Sequence similar (paralogous) protein groups

E.coli contains many families of proteins of similar sequence, reflecting paralogous groups of genes (7–9). In order to delineate the paralogous groups by sequence similarity, one must separate the multimodular proteins into units of independent origin and function to aviod artificial connections between unrelated proteins (9). Using the revised list of

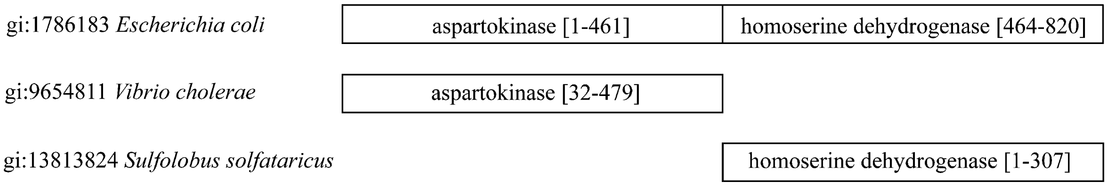
gi:1786183 *Escherichia coli*
aspartokinase [1-461]
homoserine dehydrogenase [464-820]
gi:9654811 *Vibrio cholerae*
aspartokinase [32-479]
gi:13813824 *Sulfolobus solfataricus*
homoserine dehydrogenase [1-307]

Figure 1. Identification of a fused gene product. Darwin-generated sequence alignments of *E.coli* proteins against proteins of 50 genomes were used to identify gene products with two or more separately encoded functions. The alignments with ThrA (GI:1786183) support the separation of the protein into two modules. The full-length alignment of *Vibrio cholerae* aspartokinase locates the N-terminally encoded function while the full-length alignment of the *Sulfolobus solfataricus* homoserine dehydrogenase locates the C-terminally encoded function. Alignment regions for the respective proteins are shown in brackets.

modules, pairwise protein sequence similarities were detected by Darwin (6) as detailed above. The proteins were clustered by a transitive grouping process as described by Liang *et al.* (7). The revised grouping placed 1999 protein modules in 498 protein groups. Group size ranges from 2 to 93 members. The majority of the groups have two (55%) or three (20%) members. Only 29 groups have 10 or more members. Annotation information for unknown proteins can be derived from membership in sequence-similar groups whose common function is clear.

Cellular role (MultiFun) assignments for *E.coli* gene products

The *E.coli* gene products are characterized by their cellular role according to the MultiFun classification system (3). This is a hierarchical classification system consisting of 10 main role categories that are further subdivided. The classification system has been expanded recently to include additional metabolic pathways as well as categories for the substrates of transport proteins. MultiFun is currently also used by A Systematic Annotation Package for community analysis of genomes (ASAP) (10) and in the EcoCyc database (11). We continually make cellular role assignments to our database and have at the present made over 8400 assignments representing on an average more than two assignments per gene product. Assigning multiple cell roles to a gene product is done to more completely characterize its activity in the cell.

SUMMARY

GenProtEC presents information on the functions of *E.coli* K-12 MG1655 gene products from several points of view. *E.coli* proteins as single modules have been grouped in sequence similarity. Using the power of group membership of proteins of similar function, open reading frames within any group can be assigned the general function. In addition, the presence of domains of known function within *E.coli* proteins has been determined. Domain content permits annotation of some functional information to otherwise totally unknown sequences. The rich classification of cellular roles, MultiFun, has been applied, underlining the fact that many gene products have more than one cellular role. All the data presented in GenProtEC is made easily accessible to the users through downloadable flat files in text format.

Supplementary data are available online as follows: gene page list for ThrA (http://genprotec.mbl.edu/result.php?search_field=Gene+Name&field_value=thra); multimodular *E.coli* proteins (http://genprotec.mbl.edu/prot_mod.php); groups of *E.coli* proteins (modules) related by sequence similarity (http://genprotec.mbl.edu/prot_grp.php); groups of *E.coli* proteins related by biochemistry and structure (http://genprotec.mbl.edu/bio_group.html).

ACKNOWLEDGEMENTS

We thank Amar Singh for redesigning and programming GenProtEC. Thank you to Edward A. Adelberg for providing up to date references on *E.coli* gene products. Incorporation of SCOP domains was done in collaboration with Sarah A. Teichmann.

REFERENCES

1. Riley,M. and Space,D.B. (1996) Genes and proteins of *Escherichia coli* (GenProtEc). *Nucleic Acids Res.*, **24**, 40.
2. Serres,M.H., Gopal,S., Nahum,L.A., Liang,P., Gaasterland,T. and Riley,M. (2001) A functional update of the *Escherichia coli* K-12 genome. *Genome Biol.*, **2**, RESEARCH0035.
3. Serres,M.H. and Riley,M. (2000) MultiFun, a multifunctional classification scheme for *Escherichia coli* K-12 gene products. *Microb. Comp. Genomics*, **5**, 205–222.
4. Lo,C.L., Brenner,S.E., Hubbard,T.J., Chothia,C. and Murzin,A.G. (2002) SCOP database in 2002: refinements accommodate structural genomics. *Nucleic Acids Res.*, **30**, 264–267.
5. Gough,J. and Chothia,C. (2002) SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments. *Nucleic Acids Res.*, **30**, 268–272.
6. Gonnet,G.H., Hallett,M.T., Korostensky,C. and Bernardin,L. (2000) Darwin v. 2.0: an interpreted computer language for the biosciences. *Bioinformatics*, **16**, 101–103.
7. Liang,P., Labedan,B. and Riley,M. (2002) Physiological genomics of *Escherichia coli* protein families. *Physiol. Genomics*, **9**, 15–26.
8. Liang,P. and Riley,M. (2001) A comparative genomics approach for studying ancestral proteins and evolution. *Adv. Appl. Microbiol.*, **50**, 39–72.
9. Labedan,B. and Riley,M. 1995. Gene products of *Escherichia coli*: Sequence comparisons and common ancestries. *Mol. Biol. Evol.*, **12**, 980–987.
10. Glasner,J.D., Liss,P., Plunkett,G.,III, Darling,A., Prasad,T., Rusch,M., Byrnes,A., Gilson,M., Biehl,B., Blattner,F.R. *et al.* (2003) ASAP, a systematic annotation package for community analysis of genomes. *Nucleic Acids Res.*, **31**, 147–151.
11. Karp,P.D., Riley,M., Saier,M., Paulsen,I.T., Collado-Vides,J. S., Paley,M., Pellegrini-Toole,A., Bonavides,C. and Gama-Castro,S. (2002) The EcoCyc Database. *Nucleic Acids Res.*, **30**, 56–58.